artificiality

MINDS MEETING MACHINES

Industry Updates

Mechanistic Interpretability

Memory vs Margins

Preview: How to Talk to Your Teens

Industry Updates

Mechanistic Interpretability

Memory vs Margins

Preview: How to Talk to Your Teens

Things of Note

- **OpenAI:** Management mess, GPT Store delayed, data leaks.
- **Google:** Releases Gemini, its next generation foundation model, and more.
- **Apple:** Quietly open sources MLX.
- **Amazon:** Releases Q with severe hallucinations and leaky data.
- **xAI Grok:** “Will have problems,” seeking $1B.
- **Anthropic Claude:** Doubled context window, halved hallucations.
- **Mistral:** Releases open source, Sparse Model of Experts LLM.
- **IBM, Meta, and many others:** Launched AI Alliance to advance open science in AI.
- **EU AI Act:** In effect 2025; addresses facial image scraping, emotion recognition, social scoring, training transparency.

OpenAI

What Happened

- CEO Sam Altman was fired by the board…and then Altman returned and fired the board.
- Released GPT-4 Turbo, a smaller, cheaper, and more up-to-date foundation model…and then Google found a way to access training data.
- Announced GPTs, an innovative way for individuals to create their own LLMs…but outsiders found ways to access user data…and then OpenAI shut off new user access.
- Announced the GPT Store, an app store-like commerce opportunity for GPT developers…and then delayed launch into 2024.
- Announced enterprise services…competing with its #1 customer/partner/investor, Microsoft.

Why it Matters

- Complete lack of stability at the company that is attempting to create AGI, an AI that can surpass human capabilities at all economically valuable work.
- The Altman saga means speed & scale won over caution, profits won over care. The non-profit structure that was supposed to save the world is simply a farce.
- The company's technical advances are very impressive…but the unpredictability of its testing, security, privacy, and product plans make it hard to rely on.
- The new board is (so far) 100% white men of privilege, a stunning statement from a company that intends to benefit all humanity.

Google

What Happened

- Announced Gemini, Google's next generation foundation model which is multi-model from the ground up.
- Benchmark tests show performance exceeding OpenAI's GPT…but benchmarks are only useful to a point.
- Three sizes: Ultra, Pro, Nano…but not all models are available today so can't evaluate yet.
- Demoed innovative uses within application workflows…but the demos were aspirational rather than real.
- Demoed dynamic coding of bespoke UX.

Why it Matters

- Google has been seen as trailing OpenAI (despite creating the technology behind LLMs) so releasing a leading model re-establishes Google as a player in foundation models.
- Multi-model is the future and an integrated model may outperform longer term.
- Focusing on application workflow shows the future of how companies like Google, Microsoft, and Apple might implement generative AI and fits within our obsession, *World of Workflows*.
- Dynamic coding fits within our obsession, *Dynamic Design*.
- Nano fits within our obsession, *Mobile Matters*.

Benchmarks

- Benchmarks: Useful? Yes. But within limits.
- Some users will want an AI to tackle these tasks discreetly but *many* more will want these as part of a workflow—and that workflow involves software and hardware that isn't represented here.

	Gemini Ultra	Gemini Pro	GPT-4	GPT-3.5	PaLM 2-L	Claude 2	Inflect-ion-2	Grok 1	LLAMA-2
MMLU Multiple-choice questions in 57 subjects (professional & academic) (Hendrycks et al., 2021a)	**90.04%** CoT@32*	79.13% CoT@8*	87.29% CoT@32 (via API**)	70% 5-shot	78.4% 5-shot	78.5% 5-shot CoT	79.6% 5-shot	73.0% 5-shot	68.0%***
	83.7% 5-shot	71.8% 5-shot	86.4% 5-shot (reported)						
GSM8K Grade-school math (Cobbe et al., 2021)	**94.4%** Maj1@32	86.5% Maj1@32	92.0% SFT & 5-shot CoT	57.1% 5-shot	80.0% 5-shot	88.0% 0-shot	81.4% 8-shot	62.9% 8-shot	56.8% 5-shot
MATH Math problems across 5 difficulty levels & 7 subdisciplines (Hendrycks et al., 2021b)	**53.2%** 4-shot	32.6% 4-shot	52.9% 4-shot (via API**)	34.1% 4-shot (via API**)	34.4% 4-shot	—	34.8%	23.9% 4-shot	13.5% 4-shot
			50.3% (Zheng et al., 2023)						
BIG-Bench-Hard Subset of hard BIG-bench tasks written as CoT problems (Srivastava et al., 2022)	**83.6%** 3-shot	75.0% 3-shot	83.1% 3-shot (via API**)	66.6% 3-shot (via API**)	77.7% 3-shot	—	—	—	51.2% 3-shot
HumanEval Python coding tasks (Chen et al., 2021)	**74.4%** 0-shot (IT)	67.7% 0-shot (IT)	67.0% 0-shot (reported)	48.1% 0-shot	—	70.0% 0-shot	44.5% 0-shot	63.2% 0-shot	29.9% 0-shot
Natural2Code Python code generation. (New held-out set with no leakage on web)	**74.9%** 0-shot	69.6% 0-shot	73.9% 0-shot (via API**)	62.3% 0-shot (via API**)	—	—	—	—	—
DROP Reading comprehension & arithmetic. (metric: F1-score) (Dua et al., 2019)	**82.4** Variable shots	74.1 Variable shots	80.9 3-shot (reported)	64.1 3-shot	82.0 Variable shots	—	—	—	—
HellaSwag (validation set) Common-sense multiple choice questions (Zellers et al., 2019)	87.8% 10-shot	84.7% 10-shot	**95.3%** 10-shot (reported)	85.5% 10-shot	86.8% 10-shot	—	89.0% 10-shot	—	80.0%***
WMT23 Machine translation (metric: BLEURT) (Tom et al., 2023)	**74.4** 1-shot (IT)	71.7 1-shot	73.8 1-shot (via API**)	—	72.7 1-shot	—	—	—	—

The Dawn of Dynamic Design

Apple

What Happened

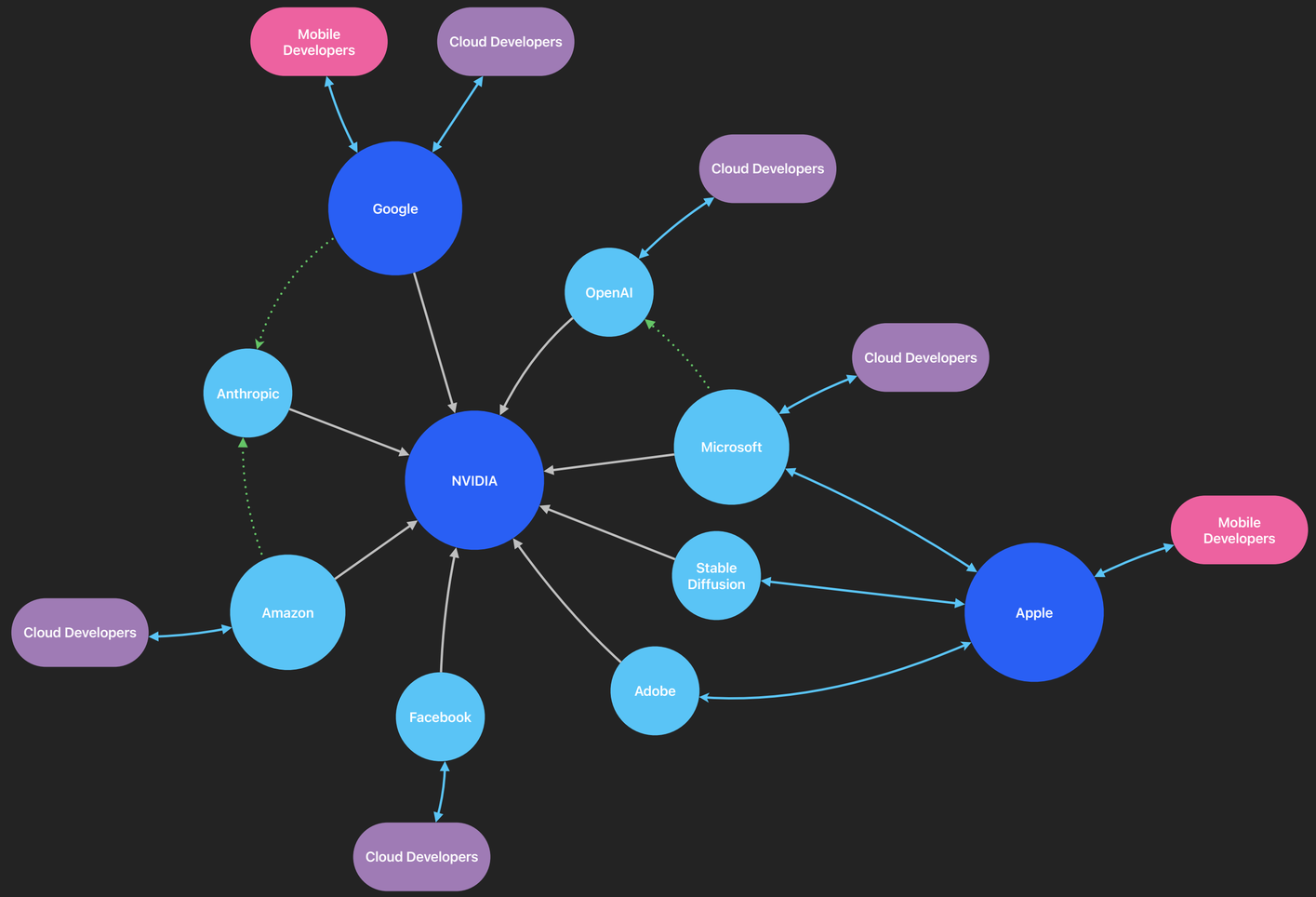
- Quietly released MLX, a developer framework for generative AI on Apple Silicon.
- Accelerates open source models including LLaMA (LLM from Facebook), Stable Diffusion (text-to-video), and Whisper (speech-to-text from OpenAI).
- Includes a unified memory model, allowing shared memory access for CPU and GPU, increasing efficiency and speed.
- Includes dynamic graph construction, eliminating slow compilations triggered by changes in function argument shapes.

Why it Matters

- Critics say Apple has been caught flat-footed by the generative AI wave and is far behind.
- This type of release shows that Apple is playing a different game than others. It isn't trying to compete with OpenAI for a new general-purpose tool like ChatGPT. It is trying to create an aiOS for itself and its developers to build on.
- Open source will continue to be an essential part of Apple's tech stack, following on BSD, Swift, WebKit.
- Apple Silicon gives the company a distinct advantage and is core to our obsession, *Mobile Matters*.

Mobile Matters

- Data center compute is expensive
- Current business models make engagement the enemy of profits
- Mobile compute is free
- Google Nano, Apple Silicon + MLX/open source sets the stage

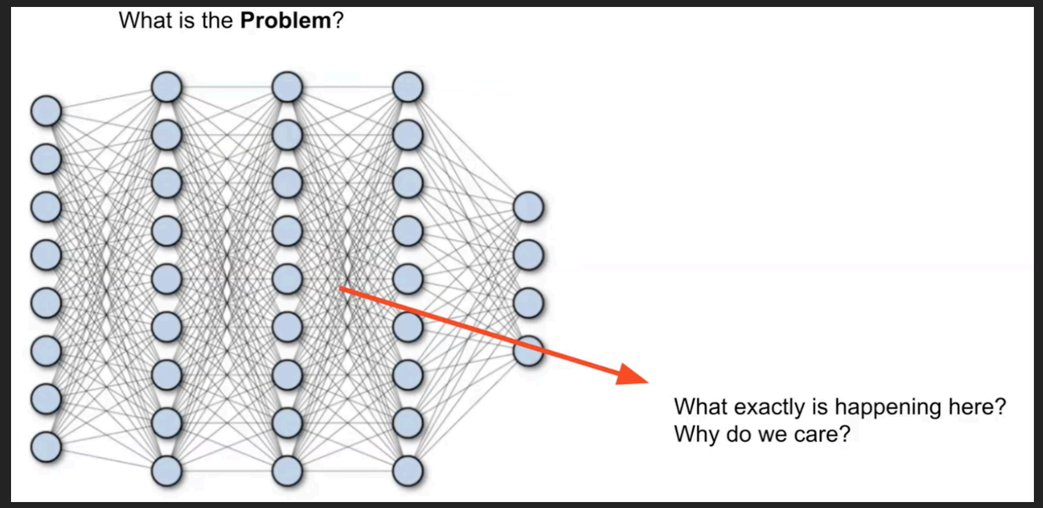


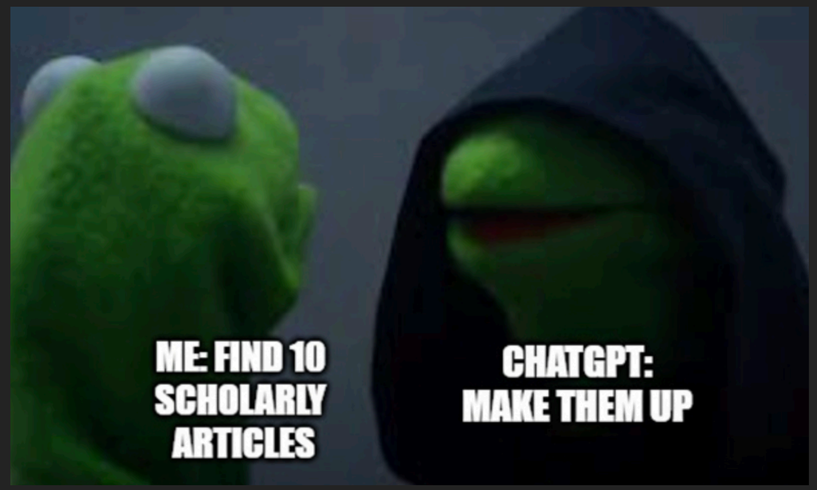
Industry Updates

Mechanistic Interpretability

Memory vs Margins

Preview: How to Talk to Your Teens





Core Hypothesis:

Models learn human, comprehensible things and can be understood.

Mechanistic Interpretability

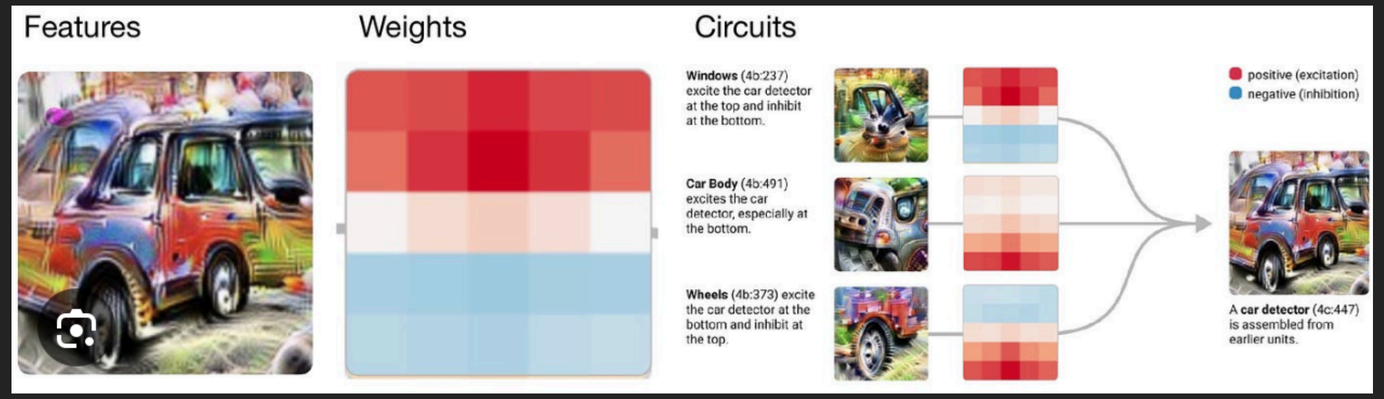
- Unraveling the “black box” of AI
- Go beyond seeing outputs and peer into the logic of its internal processes
- Progress includes the discover of features which reflect detectable patterns in data that models need to handle
- And work to identify and characterize 'circuits' within neural networks that are responsible for specific tasks or types of learning
- Increasing focus on automating parts of this interpretability research to scale up the efforts, like using AI itself to label and explain its own neural network components

Goals of Mechanistic Interpretability

- **Detailed Understanding of AI Mechanisms:** It allows for a comprehensive reverse-engineering of AI models, leading to a precise understanding of how various components contribute to the network's functionality.
- **Identification of Learning Phases:** Mechanistic interpretability aids in identifying distinct learning phases within AI models, providing insight into their complex learning dynamics.
- **Development of New Analytical Tools:** This approach facilitates the creation of new tools and metrics for analyzing AI behavior, which can be crucial for predicting and managing emergent properties in complex models.

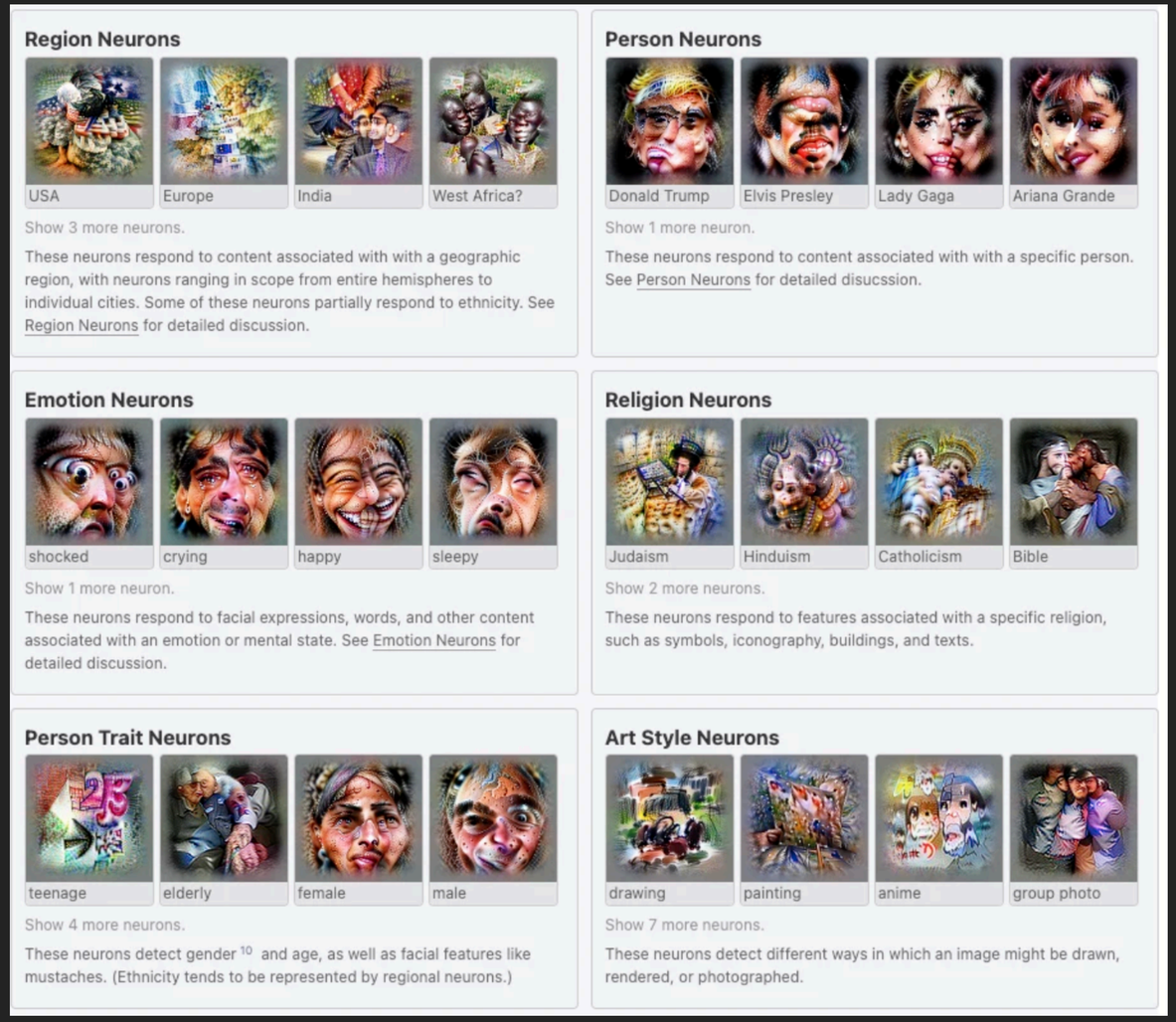
What do models look like? Features and circuits.

- Feature: some property of the input, which you can think of as some variable the model is representing inside itself.
- Circuit: algorithms the model has learned to take some features or take the input and produce more features.
- Models are feature extractors: They take inputs, and they try to find properties of them, compute them, represent them internally.



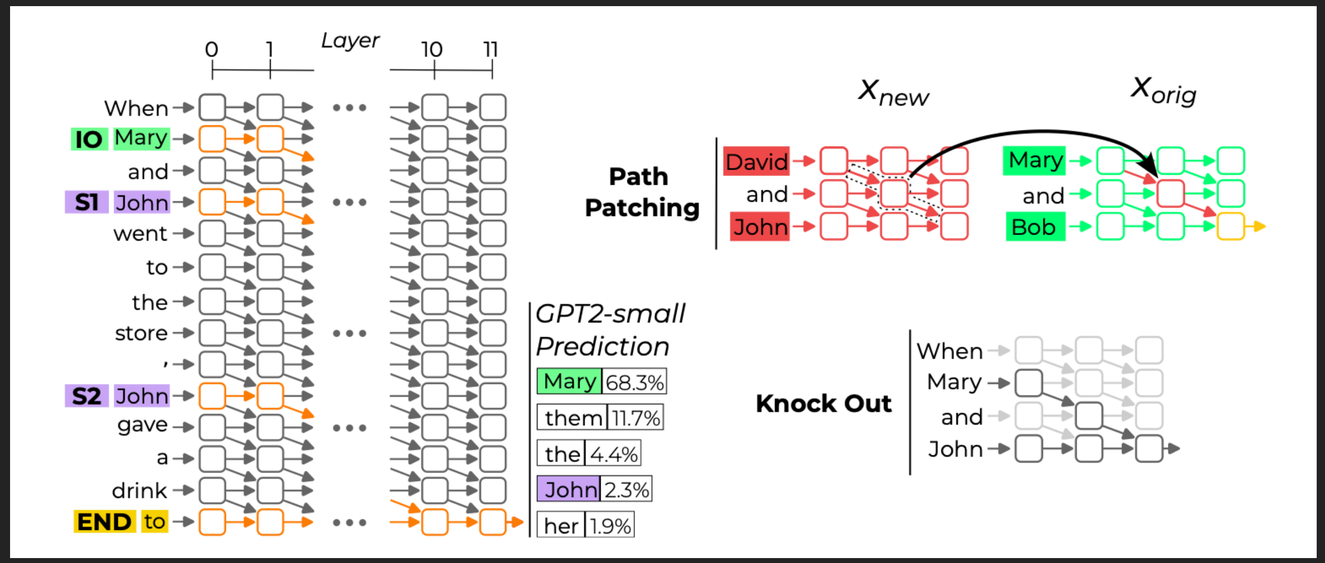
Features inside models

- This model was taking images, taking captions, and seeing if they were a good match.
- It learned features like the USA, Donald Trump, anime, the abstract notion of teenagerism.
- The Donald Trump neuron activated some things like MAGA hats and Republican politicians more than Democratic politicians.



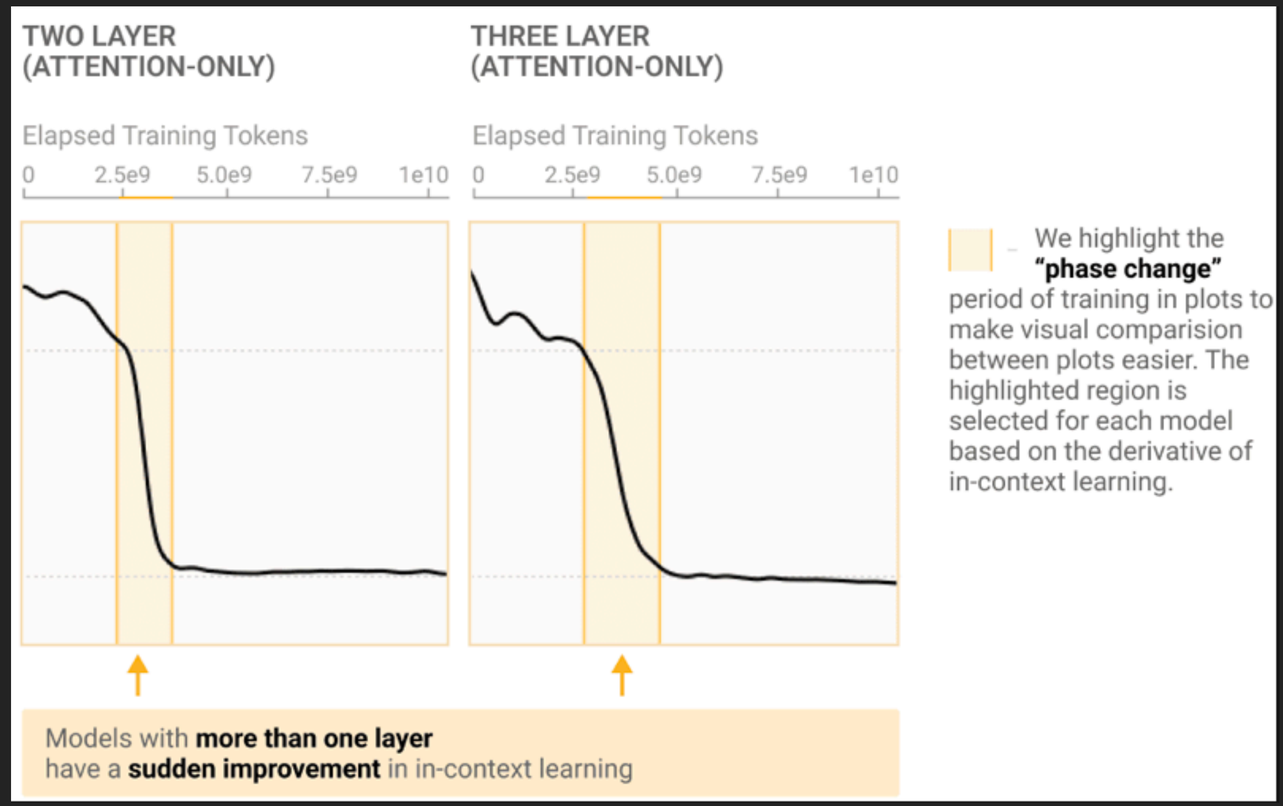
And understand the circuits, or algorithms, that the model learns

- Circuits are how the algorithm is learned by a model, how will the models “think”
- Models learn legible algorithms
- In some contexts, we can reverse-engineer them



And find emergent learning behaviors

- Language models can go beyond statistical correlations and perform in-context learning
- They can use the text they have already predicted to predict what comes next
- This is an **emergent** property - when a model is being trained it will suddenly get better at in-context learning
- This **phase change** is very intriguing and has driven researchers to develop methods to understand what the model is doing



Superposition

- Models want to represent as many features as possible but they only have so many neurons
- In superposition, neurons don't just handle one job or feature; they juggle several at the same time
- This overlapping functionality makes it hard to pinpoint what specific role each neuron plays in the network's decision-making process
- For instance a neuron which seems to respond to pictures of poetry and also to pictures of card games and dice and poker
- Disentangling this interwoven activity to understand the network's reasoning becomes a significant challenge

Progress in 2023

- Deeper understanding of how models learn to generalize rather than memorize
- Novel conceptualization of how models take advantage of scale
- Discovery of flexibility and phase shifting in algorithms dependent on structure
- Discovery of a new approach to the problem of superposition using autoencoders

Key challenges - 2024 will be a big year in mech-int

- Scaling the autoencoder—100x is compute intensive
- Scale the interpretation of the query
- Scaling in complexity is non-linear

Industry Updates

Mechanistic Interpretability

Memory vs Margins

Preview: How to Talk to Your Teens

The basic math

$ / Tokens	x	Tokens	x	Frequency	x	Users	=	Total
$0.02/1,000	x	1,000	x	10	x	10,000	=	$2,000

Note: OpenAI charges different rates for input/prompt ($0.01/1k) and output/response ($0.03k) tokens. We're using a blended $0.02 for illustration simplicity.

An email scenario

To:	Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed mattis non lorem at luctus. Aenean gravida, enim vitae vulputate mattis, orci velit ultrices odio, vel faucibus orci elit ut urna. Proin pretium est ut quam mollis accumsan. Maecenas pulvinar lorem id dignissim interdum. Cras maximus placerat metus sed euismod. Vivamus gravida, elit id malesuada tincidunt, eros risus dapibus lacus, non tincidunt velit sem ut libero. Aliquam sit amet eros cursus, porttitor felis ut, ornare erat.	100
From:	Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed mattis non lorem at luctus. Aenean gravida, enim vitae vulputate mattis, orci velit ultrices odio, vel faucibus orci elit ut urna. Proin pretium est ut quam mollis accumsan. Maecenas pulvinar lorem id dignissim interdum. Cras maximus placerat metus sed euismod. Vivamus gravida, elit id malesuada tincidunt, eros risus dapibus lacus, non tincidunt velit sem ut libero. Aliquam sit amet eros cursus, porttitor felis ut, ornare erat.	100
		200

To:	Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed mattis non lorem at luctus. Aenean gravida, enim vitae vulputate mattis, orci velit ultrices odio, vel faucibus orci elit ut urna. Proin pretium est ut quam mollis accumsan. Maecenas pulvinar lorem id dignissim interdum. Cras maximus placerat metus sed euismod. Vivamus gravida, elit id malesuada tincidunt, eros risus dapibus lacus, non tincidunt velit sem ut libero. Aliquam sit amet eros cursus, porttitor felis ut, ornare erat.	100
From:	Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed mattis non lorem at luctus. Aenean gravida, enim vitae vulputate mattis, orci velit ultrices odio, vel faucibus orci elit ut urna. Proin pretium est ut quam mollis accumsan. Maecenas pulvinar lorem id dignissim interdum. Cras maximus placerat metus sed euismod. Vivamus gravida, elit id malesuada tincidunt, eros risus dapibus lacus, non tincidunt velit sem ut libero. Aliquam sit amet eros cursus, porttitor felis ut, ornare erat.	100
To:	Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed mattis non lorem at luctus. Aenean gravida, enim vitae vulputate mattis, orci velit ultrices odio, vel faucibus orci elit ut urna. Proin pretium est ut quam mollis accumsan. Maecenas pulvinar lorem id dignissim interdum. Cras maximus placerat metus sed euismod. Vivamus gravida, elit id malesuada tincidunt, eros risus dapibus lacus, non tincidunt velit sem ut libero. Aliquam sit amet eros cursus, porttitor felis ut, ornare erat.	300
From:	Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed mattis non lorem at luctus. Aenean gravida, enim vitae vulputate mattis, orci velit ultrices odio, vel faucibus orci elit ut urna. Proin pretium est ut quam mollis accumsan. Maecenas pulvinar lorem id dignissim interdum. Cras maximus placerat metus sed euismod. Vivamus gravida, elit id malesuada tincidunt, eros risus dapibus lacus, non tincidunt velit sem ut libero. Aliquam sit amet eros cursus, porttitor felis ut, ornare erat.	100
To:	Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed mattis non lorem at luctus. Aenean gravida, enim vitae vulputate mattis, orci velit ultrices odio, vel faucibus orci elit ut urna. Proin pretium est ut quam mollis accumsan. Maecenas pulvinar lorem id dignissim interdum. Cras maximus placerat metus sed euismod. Vivamus gravida, elit id malesuada tincidunt, eros risus dapibus lacus, non tincidunt velit sem ut libero. Aliquam sit amet eros cursus, porttitor felis ut, ornare erat.	500
From:	Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed mattis non lorem at luctus. Aenean gravida, enim vitae vulputate mattis, orci velit ultrices odio, vel faucibus orci elit ut urna. Proin pretium est ut quam mollis accumsan. Maecenas pulvinar lorem id dignissim interdum. Cras maximus placerat metus sed euismod. Vivamus gravida, elit id malesuada tincidunt, eros risus dapibus lacus, non tincidunt velit sem ut libero. Aliquam sit amet eros cursus, porttitor felis ut, ornare erat.	100
		1,200

An email scenario

Memory is required for experience, but it's expensive

$ / Tokens	x	Tokens	x	Frequency	x	Users	=	Total
$0.02/1,000	x	200	x	10,000	x	10,000	=	$400,000
$0.02/1,000	x	1,200	x	10,000	x	10,000	=	$2,400,000

A chatbot scenario

A viral app could be a disaster

$ / Tokens	x	Tokens	x	Frequency	x	Users	=	Total
$0.02/1,000	x	50	x	150	x	150M	=	$22.5M
$0.02/1,000	x	200	x	150	x	150M	=	$90.0M
$0.02/1,000	x	128,000	x	150	x	150M	=	$57.6B

Mobile Matters

On device compute is essentially free

$ / Tokens	x	Tokens	x	Frequency	x	Users	=	Total
$0	x	∞	x	∞	x	∞	=	$0

Industry Updates

Mechanistic Interpretability

Memory vs Margins

Preview: How to Talk to Your Teens

How to Talk to Your Teens

- **Key topics:**
  - **The Good and the Bad of AI:** What it is, bias, education, how to use.
  - **Mental Health:** Accessing mental health support can be difficult. Generative AI while largely still an experiment, has the potential to empower young people struggling alone quietly.
  - **Social:** AI generated characters can engage in conversations, simulate various personalities, and provide an interactive experience that's both engaging and educational. For teens, this technology can be a source of entertainment, learning, and even companionship.
  - **Career:** The careers of the future will be about *using* AI not being replaced by AI.
  - **The Brain on AI:** When it comes to determining your teen's future self, AI may have the upper hand.

How to Talk to Your Teens

- **Online event:** Friday December 22 at 11am Pacific
- **Sign up for Artificiality** (www.artificiality.world)
  - Select “Free” plan with your Starbucks email address (you will be defaulted to Pro)
  - Invite sent via email
  - Recording will be available on Artificiality
- **Open to all:** Invite your partners & friends!

Artificiality Pro

- Actionable intelligence through monthly premium research briefings, including updates on industry shifts, product releases, research and science advances, and cross-industry best practices.
- Key insights and meta-research through essays and podcasts on the human impact of AI.
- Strategic support through presentations, frameworks, and how-tos.
- Advisory and coaching through 1:1 access to our founders for key individuals, teams, and task forces.
- Expert education and inspiration through in-person or virtual speaking and workshops.
- Enterprise-wide digital access to the Artificiality platform.